An Efficient Rotation and Translation Decoupled Initialization from Large Field of View Depth Images

Renato Martins, Eduardo Fernandez-Moral, Patrick Rives

An Efficient Rotation and Translation Decoupled Initialization from Large Field of View Depth Images

Renato Martins[1,2], Eduardo Fernandez-Moral[1] and Patrick Rives[1]

***Abstract*— Image and point cloud registration methods compute the relative pose between two images. Commonly used registration algorithms are iterative and rely on the assumption that the motion between the images is small. In this work, we propose a fast pose estimation technique to compute a rough estimate of large motions between depth images, which can be used as initialization to dense registration methods. The main idea is to explore the properties given by planar surfaces with co-visibility and their normals from two distinct viewpoints. We present, in two decoupled stages, the rotation and then the translation estimation, both based on the normal vectors orientation and on the depth. These two stages are efficiently computed by using low resolution depth images and without any feature extraction/matching. We also analyze the limitations and observabilty of this approach, and its relationship to ICP point-to-plane. Notably, if the rotation is observable, at least five degrees of freedom can be estimated in the worst case. To demonstrate the effectiveness of the method, we evaluate the initialization technique in a set of challenging scenarios, comprising simulated spherical images from the Sponza Atrium model benchmark and real spherical indoor sequences.**

I. Introduction

Image and point cloud registration are important problems in robotics and computer vision applications. The goal of registration techniques is to compute the motion, i.e., the relative pose from images. In special, mobile robotics applications require efficient registration algorithms, which are often iterative and assume a good initial pose initialization to converge (e.g., [1, 2, 3, 4, 5]). This paper describes a fast pose estimation technique using the normals from low resolution depth images. Surprisingly, except in [6, 7, 8], the information gathered from normal vectors has been exploited mainly to outlier rejection in point cloud registration algorithms as shown, for instance, in the ICP survey in [1] or in [9]. Because of its efficiency and large domain of convergence, this technique can be used as initialization to dense registration methods. Moreover, to further increase this convergence domain, we explore wide field of view (FOV) depth images without any feature extraction/matching. These images can be acquired in different ways such as, for instance, from stereo using a rig of perspective (e.g., [10]) or ominidirectional cameras (e.g., [11]), 3D LIDARs or with a rig of RGB-D sensors (e.g., [12]).

Our method estimates the rotation and translation sequentially in a decoupled fashion. The only assumed hypothesis, in the case of general scenes, is that the frames contains piece-wise planar regions with co-visibility (overlapping). A simplified geometric scheme is given in fig. 1. The rotation explores the properties of the overlapped planar regions using a projector decomposition of the normals. Subsequently, an over-determined set of linear equations can be set for the translation. We present a salience point selection to improve the conditioning of this system, as well as the limitations and observability conditions of the approach, showing that if the rotation estimation is successful, the translation can be estimated efficiently for at least two degrees of freedom (DOF). Moreover, we also remark that a Matlab non-optimized version of our algorithm runs at 20Hz (with Matlab 2012 in a laptop Intel Core i5-5300U CPU, 2.3 GHz and Ubuntu 14.04) and hence can be used on-line.

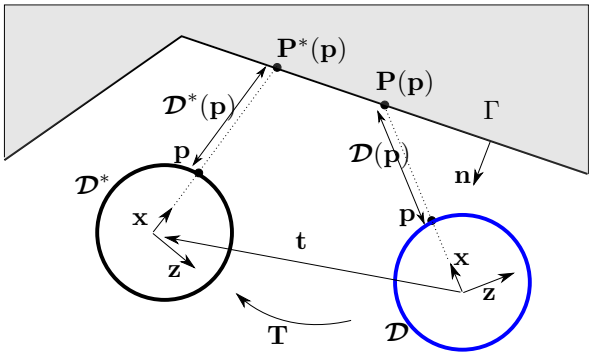

Fig. 1. Bird's-eye view schematic of two spherical frames $\mathcal{D}^*$ and $\mathcal{D}$ observing a planar region Γ. See the text in section III for notation details.

The rest of this paper is organized as follows. First, we discuss some related works in section II. A review of the spherical model and its elementary properties is given in section III. Then, we introduce the rotation initialization in section IV-A. The translation is subsequently described in section IV-B. The limitations and observability conditions of the approach are discussed in sections IV-A.1, IV-B.1 and IV-C. Experimental results are presented in section V and we conclude the paper in section VI.

II. Main Related Works

This work is directly related to the methods of [6, 7] for point-cloud registration, [8] for rotation tracking in piecewise planar environments and [12] for automatic lidar/RGB-D non-overlapping camera calibration. The approach in [6] explored the 3D-NDT transform to describe the scene using distributions of geometric features (corners, planes, lines). [7] proposed a decoupled rotation and translation estimation using the normals from two point clouds. The rotation estimation tracks the peak of the normal distributions

[1]Inria, Université Côte d'Azur, Sophia Antipolis, France.
Email: {renato-jose.martins, eduardo.fernandez-moral, patrick.rives}@inria.fr
[2]MINES ParisTech, PSL Research University, Sophia Antipolis, France.

using a decomposition similar to the one presented in this work. However, the rotation estimation requires a dominant mode to ensure the distribution mode tracking. Furthermore, their translation estimation is only valid to Manhattan-World scenes and small displacements. Similarly, [8] estimate the rotation (no translation) from a set of dominant planes in the scene. Their algorithm starts by extracting the principal orientations of the normals of the environment. The association, between the normals belonging to the modes in the current and in the reference frames, is done by considering the closest mode in a conical region, with the angle of the cone being the maximum angle allowed for the rotation. Once the association is performed, the rotation estimation is based on the same formulation presented in [12]. In [12], a rough guess of the relative rotation between the current and reference frames is provided by the user for calibrating non-overlapping RGB-D cameras. Once the association is established, an elegant modified version of Arun's algorithm of ICP point-to-point [13] is derived to find the rotation in a least square sense. It is also worth noting that both [12, 8] assume a matching of the normal vectors in some stage of their formulations. Here instead, we proceed with a different strategy and formulation. First, we do not insert a rough guess [12], not assume scenes with main directions or infinitesimal/small changes in the rotation [8] (remind that a rough relative motion is what we seek). Besides that, unlike [7, 8], we also derive a closed form for the translation and analyze the limitations and what is the expected performance of the approach in a set of scene configurations. Some other interesting works assume further hypothesis in the scene geometry, as the Manhattan World assumption in [14] for scene reconstruction and in [15] for depth registration using principal component analysis of the normal vectors.

III. Preliminaries and Spherical Representation

The example of excellence of wide FOV images is the spherical view (in the unit sphere), which is defined as those images whose FOV comprises 360 degrees in the horizontal plane. We adopt spherical depth images $\boldsymbol{\mathcal{D}} \in \mathbb{R}_+^{m\times n}$, as a basic representation, because most wide FOV images can be represented in the unit sphere $\mathbb{S}^2$ through a calibration procedure[1]. The mapping between 3D Cartesian coordinates $\mathbf{P} \in \mathbb{R}^3$ and frame pixel coordinates $\mathbf{p} \in \mathbb{P}^2$ is given by $\mathbf{P}(\mathbf{p}) = \boldsymbol{\mathcal{D}}(\mathbf{p})\Pi_S^{-1}(\mathbf{p})$, with the unit vector $\Pi_S^{-1}(\mathbf{p}) \in \mathbb{S}^2$ being the viewing direction of the 3D point $\mathbf{P}$ (see fig. 1 for the geometry of two 3D points viewed from the X sensor direction in two different frames). The relative pose between the frames is represented in the angle/axis and in the matrix form $\mathbf{T} = (\mathbf{R}, \mathbf{t}) \in \mathbb{SE}(3)$ (rotation and translation). The spherical normalization operator $\| \ \|_S$: $\mathbb{R}_+^3 \rightarrow \mathbb{S}^2$ of a 3D point is defined as

$$\|\mathbf{P}\|_S := \mathbf{P}/\|\mathbf{P}\|_2 \in \mathbb{S}^2, \text{ for } \|\mathbf{P}\|_2 \neq 0. \tag{1}$$

We introduce now the two basic geometric concepts between a rotation and two given unit vectors $\mathbf{n_1}, \mathbf{n_2} \in \mathbb{R}_+^3$. The angle Θ and orthogonal axis $\mathbf{n}_\Theta$ (perpendicular to the plane formed by the two vectors) is given by:

$$\Theta = \arccos(\mathbf{n_1^T n_2}) \text{ and } \mathbf{n}_\Theta = \|\mathbf{S}(\mathbf{n_1})\mathbf{n_2}\|_S \tag{2}$$

where $\mathbf{S}(\mathbf{n_1})$ represents the skew-symmetric matrix of the vector $\mathbf{n_1}$, such as that the cross product $\mathbf{n_1} \times \mathbf{n_2} = \mathbf{S}(\mathbf{n_1})\mathbf{n_2}$. The rotation $\mathbf{R}$ thereby establishing $\mathbf{n_1} = \mathbf{R}\mathbf{n_2}$ is

$$\mathbf{R} = \exp(\mathbf{S}(\Theta\mathbf{n}_\Theta)) \tag{3}$$

which can be computed using the well known Rodrigues' formula. For numerical stability of (2) and (3), $\mathbf{R}$ is the identity matrix if $||\Theta||_1 < 0.001$ degrees. By last, the superscript $*$ designates variables in the reference frame $\boldsymbol{\mathcal{D}}^*$.

IV. Decoupled Pose Initialization from Normals

In this section, we present an efficient way of computing not only the rotation but also the translation using normal vectors. The complete pose initialization from normals (PIN) comprises two main sequential stages: one for the rotation and one for the translation. The unique assumed hypothesis is that the scene contains co-visible planar regions. This assumption is discussed in more detail in section IV-C.

A. Rotation Initialization for General Scenes

We start describing the rotation estimation for general scenes, i.e., without the assumption of dominant directions in the normals. In presence of planar surface regions with co-visibility/overlapped (see fig. 1), the following holds: $\mathbf{n}(\mathbf{p}) = \mathbf{R}\mathbf{n}^*(\mathbf{p})$. The hypothesis of overlapped planes is quite realistic since most scenes have planar surfaces (the limitations of this hypothesis will be discussed later in section IV-C). Given the normals, the product $\arccos(\mathbf{n}^*(\mathbf{p})^T\mathbf{n}(\mathbf{p}))$ is the rotation angle around the axis $\mathbf{n}^*(\mathbf{p})\times\mathbf{n}(\mathbf{p})$. Thus, a same overlapped point have different possible rotations matrices, depending on the axis of rotation. Furthermore, a vector is invariant to a rotation around a parallel axis. Hence, an intermediary representation is used, for instance a decomposition, to find the overlapped regions.

Since any rotation can be decomposed as three instantaneous rotations around three orthogonal axes (from Euler's rotation theorem), we perform projections of all normals in three subspaces to identify the planar overlapped regions, which are rotated by the same angle in this intermediary representation. For simplicity, we select the coordinate system of the current frame $\boldsymbol{\mathcal{D}}$ to define the projection operator around each axis as

$$\begin{gathered}\text{proj}_x(\mathbf{n}) = \|(\mathbf{0}\ \mathbf{e_y}\ \mathbf{e_z})^T\mathbf{n}\|_S;\ \text{proj}_y(\mathbf{n}) = \|(\mathbf{e_x}\ \mathbf{0}\ \mathbf{e_z})^T\mathbf{n}\|_S \\ \text{proj}_z(\mathbf{n}) = \|(\mathbf{e_x}\ \mathbf{e_y}\ \mathbf{0})^T\mathbf{n}\|_S\end{gathered} \tag{4}$$

with $\mathbf{e_x} = (1\ 0\ 0)^T$, $\mathbf{e_y} = (0\ 1\ 0)^T$, $\mathbf{e_z} = (0\ 0\ 1)^T$ and $\mathbf{0} = (0\ 0\ 0)^T$. The corresponding instantaneous rotation angle of each projection $\omega_x,\ \omega_y,\ \omega_z \in [0, \pi)$ is given by the scalar products:

$$\begin{aligned}\omega_x(\mathbf{p}) &= \arccos(\text{proj}_x(\mathbf{n}^*(\mathbf{p}))^T\text{proj}_x(\mathbf{n}(\mathbf{p}))) \\ \omega_y(\mathbf{p}) &= \arccos(\text{proj}_y(\mathbf{n}^*(\mathbf{p}))^T\text{proj}_y(\mathbf{n}(\mathbf{p}))) \\ \omega_z(\mathbf{p}) &= \arccos(\text{proj}_z(\mathbf{n}^*(\mathbf{p}))^T\text{proj}_z(\mathbf{n}(\mathbf{p})))\end{aligned} \tag{5}$$

[1]Under the assumption of central cameras, i.e., all the projection rays to form the image are constrained to meet at a single point.

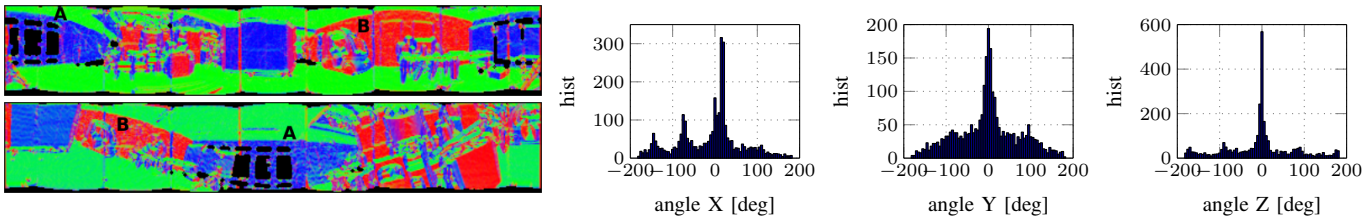

Fig. 2. Rotation estimation example for two real frames with rotation $\boldsymbol{\omega} = (21\ 175\ 0)^T$. The first column depicts the normal vectors (encoded by colors) of the frames. For visualization of the motion between the frames, we included the labels **A** and **B** in corresponding regions of the frames. The distributions of each projected angle are shown at right. Our formulation gives an estimated rotation of around $\hat{\boldsymbol{\omega}} = (20\ 0\ 0)^T$ degrees (the mode of each distribution). This example depicts a successful estimation in X, but fails to estimate a rotation of 175 degrees in Y (the overlapping property is not fulfilled). See the text of sections IV-C and V for details.

In the same way, the sign of each angle obeys the sign of the projections cross product:

$$s_i = \text{sign}(\mathbf{e}_i^T \text{proj}_i(\mathbf{n}^*(\mathbf{p})) \times \text{proj}_i(\mathbf{n}(\mathbf{p}))) \tag{6}$$

for $i = \{x, y, z\}$ as for the angles in eq. (5). Assuming that the scene contains overlapping planar regions, a rotation estimate can be obtained from the projection angles of all pixels using (4), (5) and (6). These angles can be seen as three distributions and the property we explore to extract the points with co-visibility is that overlapped planes of a same surface are rotated by the same projected angles. For instance, this is performed by finding the sub-set of pixels $\mathbf{p}^+$ belonging to the peaks of the three distributions simultaneously (e.g., the peaks in the distributions of fig. 2). With the sub-set of pixels $\mathbf{p}^+$ (inliers points), one can find the median angle of each projection:

$$\hat{\omega}_i = \text{median}(s_i(\mathbf{p}^+)\omega_i(\mathbf{p}^+)) \tag{7}$$

with $i = \{x, y, z\}$. Then, the rotation, in the axis/angle form, is given by $\boldsymbol{\omega} = (\hat{\omega}_x\ \hat{\omega}_y\ \hat{\omega}_z)^T$ and the equivalent rotation matrix is recovered by the exponential mapping $\hat{\mathbf{R}} = \exp(\mathbf{S}(\boldsymbol{\omega}))$. This algorithm is much more efficient than an ICP point-to-plane and has similar accuracy and convergence domain.

***Rotation from Normals' Mode Tracking**:* An interesting particular scenario is of scenes with normals in a dominant direction (e.g., scenes obeying the Manhattan World assumption). In this case, we can track the modes of the normals, similarly to [8, 7], by simply computing two distributions of the projections in (4) at the reference and current frames:

$$\begin{aligned} \omega_i(\mathbf{p}) &= \arccos(\mathbf{e}_i^T \text{proj}_i(\mathbf{n}(\mathbf{p}))) \\ \omega_i^*(\mathbf{p}) &= \arccos(\mathbf{e}_i^T \text{proj}_i(\mathbf{n}^*(\mathbf{p}))) \end{aligned} \tag{8}$$

And then, the three rotation angles, as in (7), are:

$$\hat{\omega}_i = \text{median}(s_i(\mathbf{p}^+)\omega_i(\mathbf{p}^+)) - \text{median}(s_i^*(\mathbf{p}^+)\omega_i^*(\mathbf{p}^+)) \tag{9}$$

with $i = \{x, y, z\}$. Considering spherical images and small translations, this case allows estimating any rotation, even for frames without surface overlapping.

*1) **Rotation Observability for the General Case**:* Let's suppose firstly that overlapped regions are given, i.e., starting from a set of inlier pixels $\mathbf{p}^+$. We want to find the rotation that minimizes the cost (for simplicity using the ℓ^2 norm):

$$\min_{\boldsymbol{\omega}} \sum_{\mathbf{p} \in \mathbf{p}^+} \frac{1}{2} ||\mathbf{e}_\mathbf{n}(\mathbf{p}, \boldsymbol{\omega})||_2^2 \tag{10}$$

where $\mathbf{e}_\mathbf{n}(\mathbf{p}, \boldsymbol{\omega})$ is the error between corresponding normal vectors observed from two distinct frames (as discussed in section IV-A),

$$\mathbf{e}_\mathbf{n}(\mathbf{p}, \boldsymbol{\omega}) = \mathbf{n}(\mathbf{p}) - \hat{\mathbf{R}}(\boldsymbol{\omega})\mathbf{n}^*(\mathbf{p}) \tag{11}$$

and $\hat{\mathbf{R}}(\boldsymbol{\omega})$ is the rotation as in eq. (3). We can proceed to a simplified local formulation for the observability, i.e., $\hat{\mathbf{e}}_\mathbf{n}(\mathbf{p}, \boldsymbol{\omega}) = \mathbf{n}(\mathbf{p}) - \hat{\mathbf{R}}\mathbf{R}(\boldsymbol{\omega})\mathbf{n}^*(\mathbf{p})$. The first order optimality condition of (10) depends of a linear approximation of the error $\hat{\mathbf{e}}_\mathbf{n}(\mathbf{p}, \boldsymbol{\omega})$ around $\hat{\mathbf{R}}$, i.e., from the Jacobian of (11):

$$\mathbf{J}(\mathbf{0}) = \hat{\mathbf{R}} \frac{\partial(\mathbf{R}(\boldsymbol{\omega})\mathbf{n}^*(\mathbf{p}))}{\partial \boldsymbol{\omega}}\bigg|_{\boldsymbol{\omega}=\mathbf{0}} = \hat{\mathbf{R}}\mathbf{S}(\mathbf{n}^*(\mathbf{p})) \in \mathbb{R}^{(3\times 3)}. \tag{12}$$

Therefore, the rotation is locally observable if the Fisher Information Matrix $\mathbf{J}(\mathbf{0})^T\mathbf{J}(\mathbf{0})$ is invertible. We can verify that given the normals at two points $\mathbf{n}^*(\mathbf{p_1})$, $\mathbf{n}^*(\mathbf{p_2})$ then $\mathbf{J}(\mathbf{0})^T\mathbf{J}(\mathbf{0}) = -\left(\mathbf{S}^2(\mathbf{n}^*(\mathbf{p_1})) + \mathbf{S}^2(\mathbf{n}^*(\mathbf{p_2}))\right)$, which is of full rank if $\mathbf{n}^*(\mathbf{p_1})$, $\mathbf{n}^*(\mathbf{p_2})$ are not parallel (co-directional). This local condition, that there is at least two planes with linearly independent normal vectors, is generally fulfilled in indoor scenarios by having often the floor, ceiling and walls not in the same direction. Furthermore, this local condition is also global by developing the cost (10) as in [12, 8]. Therefore, supposing that overlapped regions are given, the rotation is observable if the scene has, at least, two planes in linearly independent directions.

B. Translation Initialization

We proceed to the translation estimation in this section. After applying the rotation to the reference frame (also known as "derotation" process [16]), the updated overlapped surfaces for the translation is done by checking the angle between the normals. At this time, the set of overlapped pixels $\mathbf{p}^+$ are the pixels in both frames with similar normals ($\mathbf{n}(\mathbf{p}) \approx \mathbf{n}^*(\mathbf{p})$), i.e.,

$$\mathbf{p} \in \mathbf{p}^+ \text{ if } ||\arccos(\mathbf{n}^{*T}(\mathbf{p})\mathbf{n}(\mathbf{p}))||_1 < \epsilon_1. \tag{13}$$

where ϵ_1 is the maximum allowed angle between the normals. Hence, the pixels considered to be overlapped follows the same plane equation Γ. The plane equation for the 3D point in the pixel $\mathbf{p}$ of the current image is given by

$$\Gamma : \mathbf{n}^T(\mathbf{p})\mathbf{P}(\mathbf{p}) + d = 0 \Rightarrow \mathbf{n}^T(\mathbf{p})\left(\boldsymbol{\mathcal{D}}(\mathbf{p})\Pi_S^{-1}(\mathbf{p})\right) + d = 0 \tag{14}$$

with $\Pi_S^{-1}(\mathbf{p})$ the viewing direction in the unit sphere. Denoting the residual rotation $\mathbf{R}(\mathbf{p})$ for each pixel such as $\mathbf{n}^T(\mathbf{p}) = \mathbf{R}(\mathbf{p})\mathbf{n}^{*T}(\mathbf{p})$, the same plane viewed from the reference depth image in the direction $\Pi_S^{-1}(\mathbf{p})$ (as depicted in fig. 1) is therefore:

$$\Gamma : \mathbf{n}^T(\mathbf{p}) \left(\mathbf{R}(\mathbf{p})\boldsymbol{\mathcal{D}}^*(\mathbf{p})\Pi_S^{-1}(\mathbf{p}) + \mathbf{t}\right) + d = 0 \qquad (15)$$

Subtracting the left side of eq. (14) and (15), the relationship between the normal vector, depth, viewing direction and the translation (for a pixel $\mathbf{p} \in \mathbf{p}^+$) is

$$\mathbf{n}^T(\mathbf{p})\mathbf{t} = \mathbf{n}^T(\mathbf{p}) \left(\Pi_S^{-1}(\mathbf{p})\boldsymbol{\mathcal{D}}(\mathbf{p}) - \mathbf{R}(\mathbf{p})\Pi_S^{-1}(\mathbf{p})\boldsymbol{\mathcal{D}}^*(\mathbf{p})\right). \qquad (16)$$

Note that eq. (16) cannot be simplified since the scalar product $\mathbf{n}^T\mathbf{t} = \mathbf{n}^T(\mathbf{P} - \mathbf{R}\mathbf{P}^*)$ has $\mathbf{t} = \mathbf{P} - \mathbf{R}\mathbf{P}^*$ only when the translation is parallel to the normal of the plane Γ. For efficiency, the residual rotation in (16) is calculated for each pixel $\mathbf{p}$ using an approximation of eq. (3):

$$\mathbf{R}(\mathbf{p}) = \mathbf{I}_{(3\times3)} + \Theta(\mathbf{p})\mathbf{S}\left(\mathbf{n}_\Theta(\mathbf{p})\right) \qquad (17)$$

where the angle is $\Theta(\mathbf{p}) = \arccos(\mathbf{n}^{*\mathbf{T}}(\mathbf{p})\mathbf{n}(\mathbf{p}))$ and the axis $\mathbf{n}_\Theta(\mathbf{p})$ is the orthonormal vector to $\mathbf{n}^*(\mathbf{p})$ and $\mathbf{n}(\mathbf{p})$ using eq. (2). In ideal conditions, i.e., depth and normals without noise and perfect rotation estimate in section IV-A, the residual per pixel rotation $\mathbf{R}(\mathbf{p})$ is the identity matrix.

Some remarks can be drawn from equation (16): *i)* points with normals orthogonal to the motion do not contribute to the estimation ($\mathbf{n}^T\mathbf{t} = 0$ independently of $||\mathbf{t}||_2$) and; *ii)* a point with view direction orthogonal to the normal is ill-conditioned, i.e., $\mathbf{n}^T\Pi_S^{-1} \approx 0$ and consequently $||\boldsymbol{\mathcal{D}} - \boldsymbol{\mathcal{D}}^*||_1$ is unbounded. Thus, for avoiding outliers in the system (16), these points whose angle between the normal and view direction is almost orthogonal, should not be considered. If the system (16) is well-conditioned, it is efficiently solved using a robust M-estimator with, for instance, the Huber's loss function [17]. The task of finding a conditioned system is discussed in the next section.

1) ***Translation Observability and Conditioning****:* Consider the left side of the system in eq. (16) for all pixels belonging to the set $\mathbf{p}^+ = \{\mathbf{p_1}, \ldots, \mathbf{p_n}\}$. This system have a unique solution if the matrix $\mathbf{N} = [\mathbf{n}(\mathbf{p_1})\ \mathbf{n}(\mathbf{p_2})\ \ldots\ \mathbf{n}(\mathbf{p_n})]^T$ is of rank three, i.e., given at least three points from three different planes with linear independent orientations. Of course when noise is present in the normals, $\mathbf{N}$ has almost surely rank three, but then the solution of eq. (16) is merely an artifact produced by the noise.

Our goal is to reduce the conditioning of the matrix $\mathbf{N}$, i.e., the ratio of its maximum and minimum eigenvalues. We proceed, in a first moment, following the works of [10, 18] to select the 50% salient measurements of $\mathbf{N}$ that best constraints each DOF of the system. This is done by ordering the lines of $\mathbf{N}$ such that the conditioning of the subset of equations is as close to one as possible. This conditioning also gives a measure of the normals distribution in the sphere. We use the measure of the conditioning of the subset of salient lines $\mathbf{N_s}$ as an observability index. If the conditioning of $\mathrm{cond}(\mathbf{N_s}^T\mathbf{N_s}) > e_2$, the system in (16) is said to have an "ill-conditioned geometry" and we proceed to a dimension reduction. A Gaussian-Jordan elimination with partial pivoting is then used to find the column space of $\mathbf{N_s}$ and the translation estimation is done using the robust M-estimator for the two remaining DOF that are well conditioned.

C. Planar Overlapping Assumption

In this section, we discuss what are the conditions to obtain a good pose initialization and the limits of our approach in the case of general scenes. It is natural that the observability of the initialization depends on the scene geometry, i.e., in the size of the planes, their symmetry and their orientation. As stated in section IV-A.1, the rotation observability condition, that at least two planes have linearly independent normal vectors, is generally fulfilled for most scenes. The observability then remains mainly in how to extract the overlapped regions, which depends directly on the scene symmetry. The property we explore to extract the overlapped regions, presented in section IV-A in the case of general scenes, is that planes with co-visibility are rotated by the same angle. The angles are then represented as distributions and we select the peak (the mode) as being the one corresponding to the right overlapped points. The distributions, however, can have many modes in presence of geometry symmetry and the peak corresponding to the real rotation can be under-represented. Some classical examples are symmetric spaces, e.g., the sphere for any rotation $||\boldsymbol{\omega}||_2 > 0$, the cylinder with $||\boldsymbol{\omega}||_2 > 0$ around the cylinder axis or the cube with $||\boldsymbol{\omega}||_2 > \pi/4$. Other examples are described in [18]. In these cases, the distribution becomes multimodal, where one of the modes corresponds to the real rotation. Hence, this states that the rotation is not observable in general. For scenes with symmetry around a defined axis, the maximum observable angle is half the period of the symmetry. Our observability index concerning the rotation will be then related to the number of "similar" modes in the projected angle distributions.

V. Intialization Results and Discussion

In this section, we evaluate the pose initialization (PIN) in indoor simulated and real spherical sequences with challenging conditions, i.e., in environments with corridor-like scenes, large rotations and translations. We start presenting the parameters' tuning used in all experiments and then the accuracy and observability of the method. Finally, we show some results of dense RGB-D registration experiments with and without the initialization.

A. Implementation and Parameter Tuning

The normal vector in a point is computed using the central gradient, i.e., taking the left, right, top and down point neighbors. This normal computation is very efficient because the point order is directly given in our framework, still other normal vector algorithms could be explored, as for instance, the ones presented in [19]. The initialization method also admits low resolution depth images. The advantages of using low resolution depth images are twofold. First, it maintains

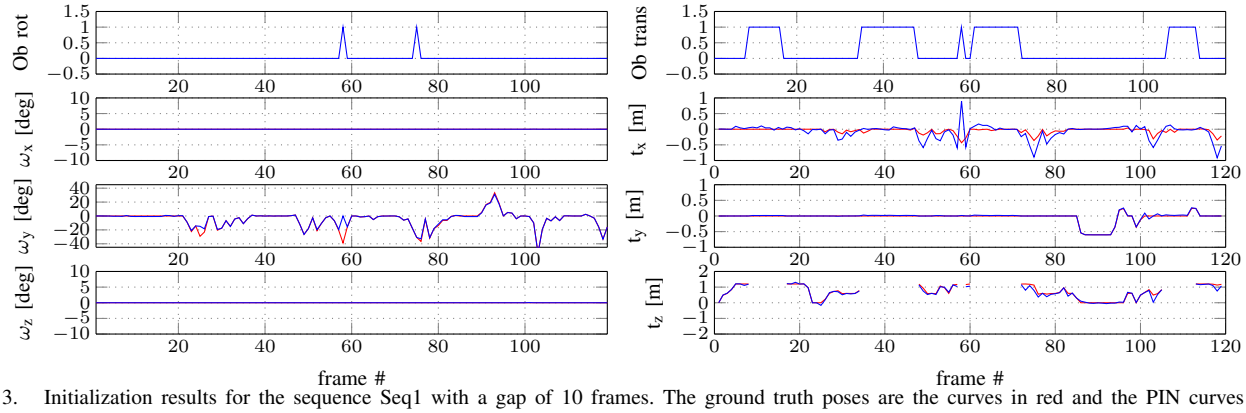

Fig. 3. Initialization results for the sequence Seq1 with a gap of 10 frames. The ground truth poses are the curves in red and the PIN curves are in blue. The graphics in the first column correspond to the rotation and the second column to the translation. The first graphic at each column depicts the observability index. The rotation index is one if the distribution is multimodal. The translation observability index is set to one, when the conditioning of the linear system for the translation is bigger than ϵ_2 = 10. See the text for details.

the efficiency of the algorithm, since a reduced number of operations are performed. Second, the central gradient is more robust to noise in the down-sampled depth. We employed a Gaussian pyramid of depth four and used the lowest resolution depth images in the experiments.

The sampling of the projected angle distributions (resolution of the histograms) was of 5 degrees to define inlier pixels. In the translation stage, points are considered to be overlapped if the angle between the normals (13) is of $\epsilon_1 = 10$ degrees. Finally, the ratio of salient pixels in IV-B.1 was set to 50% and the maximum accepted conditioning of the system without dimension reduction was heuristically set to $\epsilon_2 = 10$.

B. Pose Initialization Results

One example of the rotation initialization distributions, for a real indoor scene, is displayed in fig. 2. The procedure gives a fair estimate of the rotation of around 20 degrees in X, as can be seen in the distributions at right. Each distribution corresponds to the projected angle of rotation ω_x, ω_y and ω_z. However, this example also contains an expected failure (predicted in section IV-C), where a rotation of around 175 degrees avoids any plane overlapping around the vertical axis. Only the floor and ceiling planes are overlapped between the views.

We start using spherical depth images from the Sponza Atrium dataset, which is composed of corridors and open indoor areas. The inter-frame motion in these images are of around 0.1 meters and rotation of up to 15 degrees/frame. For checking the performance of the initialization with large motions, we have sampled the sequences with different gaps: 3, 5, 10, 15 and 20 frames – e.g. a gap of 20 frames corresponds to calculate the initialization between the image pairs (1,21), (21,41), ..., $(i, i + \text{gap})$ which results in translations of up to 2.1 meters and rotations of up to 70 degrees. The relative pose errors (RPE) for all the experiments are shown in table I as: the mean absolute error, the absolute standard deviation and the median absolute error.

Fig. 4. Omnidirectional RGB-D acquisition device with 8 Asus Xtion Pro Live (Asus XPL) sensors mounted vertically in a radial configuration (left plot) and respective point cloud of one of the offices used in the real sequence (right plot).

Spherical Simulated Sequence: Starting with the rotation, the approach proved to be robust to translations in amount of rotations of up to 2.1 meters in the sequence with gap of 20 frames. The rotation is fairly estimated in more than 99% of the cases with a gap of 10 frames with a mean absolute error of 0.8 degrees. We show the results for the gap of 10 images in fig 3 for both rotation and translation. The method failed in 10% (6/59) of cases for the experiment with a gap of 20 frames. These cases happened when the reference frame was almost completely occluded in the current frame (e.g., 90 degrees corners) and because of the scene symmetry. These failure cases were expected to happen as discussed in section IV-C and were detected by the observability index, which is displayed in the first plot of fig. 3. The translation estimation is done after warping the reference depth image using the rotation. As stated in section IV-B.1, the DOF for which the FIM is ill-conditioned cannot be accurately estimated using this formulation, some examples are depicted in fig. 3, where the t_z component could not be estimated in the frames acquired in corridors-like scenes. These cases were also predicted in section IV-B.1 and the translation index show the detected cases in the first plot of the right column. Additionally, an ICP technique in these frames with "ill-conditioned geometry" is also likely to fail to converge.

Spherical Indoor Real Sequences: We performed similar experiments using real spherical images. These real sequences were acquired in the hall and offices of the Inria building using the indoor omnidirectional RGB-D acquisition rig mounted on an holonomic mobile robot (see fig. 4). The first real sequence (seq2) is composed of 430 spherical

TABLE I

ROTATION AND TRANSLATION PIN ERRORS FOR ALL SEQUENCES – MEAN ABSOLUTE RELATIVE POSE ERROR (RPE), ABSOLUTE STANDARD DEVIATION AND ABSOLUTE MEDIAN ERROR.

	Rot RPE [deg]			Trans RPE [m]		
	$\overline{\|\|\boldsymbol{\omega}\|\|}$	std$\|\|\boldsymbol{\omega}\|\|$	med$\|\|\boldsymbol{\omega}\|\|$	$\overline{\|\|\mathbf{t}\|\|}$	std$\|\|\mathbf{t}\|\|$	med$\|\|\mathbf{t}\|\|$
Seq1 gap 5	0.25	0.39	0.06	0.11	0.18	0.04
Seq1 gap 10	0.81	3.79	0.07	0.10	0.13	0.05
Seq1 gap 15	3.02	12.58	0.09	0.32	0.61	0.12
Seq1 gap 20	6.35	21.3	0.11	0.41	0.69	0.12
Seq2 gap 3	5.12	12.34	1.18	0.30	0.33	0.25
Seq3 gap 20	7.04	19.21	1.46	0.35	0.39	0.20

TABLE II

DENSE RGB-D REGISTRATION FAILURE RATE WITH AND WITHOUT PIN.

	Seq1 gap 20	*Seq2 gap 3*
RGB-D	47/59	51/143
RGB-D+PIN	23/59	36/143

images with fast Y axis turns of up to 25 degrees between consecutive frames and with translations of around 0.15 meters. Conversely, the second real sequence (seq3) is acquired with moderate rotations of up to 5 degrees around the Y axis. To emulate large displacements, we selected a gap of 3 and 20 frames respectively. The rotation estimation was successful in 90% of cases having motions of up to 70 degrees for seq2 and up to 50 degrees in seq3. Rotations of up to 69 degrees between the real indoor frames were successfully estimated. The translation estimate is however three times more sensitive to the noise than in the simulated experiments, as shown in table I.

Finally, we use the estimated initialization within a dense RGB-D registration method. The formulation of [20] is selected with a scaling factor between the ICP and RGB costs as $\lambda = \text{median}(\boldsymbol{\mathcal{I}})/\,\text{median}(\boldsymbol{\mathcal{D}})$, where $\boldsymbol{\mathcal{I}} \in [0,\ 255]^{m\times n}$ is the intensity image. The maximum number of iterations per pyramid level is set to 20 and the method is considered to achieve convergence if the error of the final pose for the rotation and translation is smaller than 7 degrees and 0.1 meters respectively. The failure rate with (RGB-D+PIN) and without the initialization for the most challenging simulated sequence (seq1 with gap of 20 frames) and real sequence seq2 are given in table II. As expected, the initialization ensured the convergence specially in the image pairs with large rotations (bigger than 25 degrees) for the real seq2 and in the frame pairs of the simulated sequence (seq1) where the large inter-frame translation of 2.1 meters could be observed.

VI. CONCLUSIONS

This paper described a non-iterative pose estimation technique using the surface normal vectors of wide FOV depth images. The rotation and translation are computed in a decoupled way by exploring the properties of the normals of piece-wise planar scenes. First, the rotation is developed using the overlapping property of the normal vectors between two views. This is performed thanks to a decomposition of the normals in a general orthogonal coordinate system. The translation is then directly derived from the rotation as a linear system of equations. We present some techniques to improve the conditioning of this system and a discussion about our assumptions and the limits of the method. Finally, an experimental validation is performed with simulated and real spherical sequences. It is worth noting that the method does not assume Manhattan-World like scenes, small motions or any feature extraction/matching. By last, we enforce the efficiency of the initialization computation. The algorithm runs, in a Matlab non-optimized code, with around 0.02 seconds for the rotation and 0.03 seconds for the translation with Matlab 2012 in a laptop Intel Core i5-5300U CPU, 2.3 GHz and Ubuntu 14.04.

In our future research, we plan to study more appropriate metrics than the mode to match the distributions and to test our algorithm in outdoor scenes using 3D LIDAR sequences. Another future research direction is to explore other information sources, as intensity/color, if one is using RGB-D sensors.

ACKNOWLEDGMENTS

The authors thank Paolo Salaris for the proof reading of the manuscript, and the reviewers for their thoughtful comments. This work was funded by CNPq of Brazil under contract number 216026/2013-0.

REFERENCES

[1] F. Pomerleau, F. Colas, and R. Siegwart, "A review of point cloud registration algorithms for mobile robotics," *Found. Trends Robot*, vol. 4, no. 1, 2015.
[2] H. Hadj-Abdelkader, E. Malis, and P. Rives, "Spherical image processing for accurate visual odometry with omnidirectional cameras," in *IEEE OMNIVIS*, 2008.
[3] J. Kelly and G. Sukhatme, "Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration," *IJRR*, vol. 30, no. 1, 2011.
[4] R. Martins, E. Fernandez-Moral, and P. Rives, "Adaptive direct RGB-D registration and mapping for large motions," in *ACCV*, 2016.
[5] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the Mars exploration rovers," *Journal of Field Robotics*, vol. 24, no. 3, 2007.
[6] T. Stoyanov, M. Magnusson, H. Andreasson, and A. Lilienthal, "Fast and accurate scan registration through minimization of the distance between compact 3D NDT representations," *IJRR*, vol. 31, no. 12, 2012.
[7] Y. Ma, Y. Guo, J. Zhao, M. Lu, J. Zhang, and J. Wan, "Fast and accurate registration of structured point clouds with small overlaps," in *IEEE CVPR Workshops*, 2016.
[8] Y. Zhou, L. Kneip, and H. Li, "Real time rotation estimation for dense depth sensors in piece-wise planar environments," in *IEEE IROS*, 2016.
[9] J. Serafin and G. Grisetti, "NICP: dense normal based point cloud registration," in *IEEE IROS*, 2015.
[10] M. Meilland, A. Comport, and P. Rives, "Dense omnidirectional RGB-D mapping of large-scale outdoor environments for real-time localization and autonomous navigation," *JFR*, vol. 32, no. 4, 2015.
[11] M. Schönbein and A. Geiger, "Omnidirectional 3D reconstruction in augmented Manhattan worlds," in *IEEE IROS*, 2014.
[12] E. Fernandez-Moral, J. Gonzalez-Jimenez, P. Rives, and V. Arevalo, "Extrinsic calibration of a set of range cameras in 5 seconds without pattern," in *IEEE IROS*, 2014.
[13] K. Arun, T. Huang, and S. Blostein, "Least-squares fitting of two 3-D point sets," *IEEE Trans. on PAMI*, vol. 19, no. 5, 1987.
[14] H. Kim and A. Hilton, "Planar urban scene reconstruction from spherical images using facade alignment," in *IEEE IVMSP*, 2013.
[15] Y. Zhou, L. Kneip, C. Rodriguez, and H. Li, "Divide and conquer: Effcient density-based tracking of 3D sensors in Manhattan worlds," in *ACCV*, 2016.
[16] P. Corke and R. Mahony, "Sensing and control on the sphere," in *Robotics Research : The 14th International Symposium ISSR*, 2009.
[17] Z. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," Inria, Tech. Rep. 2676, 1995.
[18] N. Gelfand, L. Ikemoto, S. Rusinkiewicz, and M. Levoy, "Geometrically stable sampling for the ICP algorithm," in *3DIM*, 2003.
[19] H. Badino, D. Huber, Y. Park, and T. Kanade, "Fast and accurate computation of surface normals from range images," in *IEEE ICRA*, 2011.
[20] T. Tykkala, C. Audras, and A. Comport, "Direct iterative closest point for real-time visual odometry," in *ICCV Workshops*, 2011.